IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Towards Effective Pattern Recognition Based on Enhanced Weighted K-mean Clustering Algorithm for Groundwater Resource Planning in Point Cloud

ATIF RIZWAN[1], NAEEM IQBAL[2],ANAM NAWAZ KHAN[3] ,RASHID AHMAD [4]AND DO HYEUN KIM.[5,*],

[1,2,3,5]Department of Computer Engineering, Jeju National University, Jejusi 63243, Jeju Special Self-Governing Province, Republic of Korea;
[4]Department of Computer Science, COMSATS University Islamabad, Attock Campus 43600, Pakistan;

Corresponding author: Do Hyeun Kim (e-mail: kimdh@jejunu.ac.kr)

This research was supported by Energy Cloud RD Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Science, ICT (2019M3F2A1073387), and this work is supported by the Korea Agency for Infrastructure Technology Advancement(KAIA) grant funded by the Ministry of Land, Infrastructure and Transport (Grant 20DCRU-B158151-01). Any correspondence related to this paper should be addressed to Dohyeun Kim.

ABSTRACT Recently groundwater scarcity has accelerated drilling operations worldwide as drilled boreholes are substantial for replenishing the needs of safe drinking water and achieving long-term sustainable development goals. However, the quest for achieving optimal drilling efficiency is ever continued. This paper aims to provide valuable insights into borehole drilling data utilizing the potential of advanced analytics by employing several enhanced cluster analysis techniques to propel drilling efficiency optimization and knowledge discovery. The study proposed an L2-weighted K-mean clustering algorithm in which the mean is computed from transformed weighted feature space. To verify the effectiveness of our proposed L2-weighted K-mean algorithm, we performed a comparative analysis of the proposed work with traditional clustering algorithms to estimate the digging time and depth for different soil materials and land layers. The proposed clustering scheme is evaluated widely used evaluation metrics such as Dunn Index, Davies–Bouldin index (DBI), Silhouette coefficient (SC), and Calinski–Harabaz Index (CHI). The study results highlight the significance of the proposed clustering algorithm as it achieved better clustering results than conventional clustering approaches. Moreover, for facilitation of subsequent learning, achievement of reliable classification, and generalization, we performed feature extraction based on the time interval of the drilling process according to soil material and land layer. We formulated the solution by grouping the extracted features into six different blocks to achieve our desired objective. Each block corresponds to various characteristics of soil materials and land layers. Extracted features are examined and visualized in point cloud space to analyze the water level patterns, depth, and days required to complete the drilling operations.

INDEX TERMS data analysis; features extraction; unsupervised learning; machine learning; strategic planning and management;

I. INTRODUCTION

A. BACKGROUND

Humans need fast and easy access to safe and potable drinking water to maintain a healthy and quality life. Over the past few decades, groundwater acquisition through drilled borewells has provided access to scarce and precious groundwater resources. Since the evolution of the world, groundwater has played a significant role in the determination of origin and fate of all living beings [1]. The accessibility of groundwater has become a driving force for agriculture and economic productivity and has increased it to manifolds. Unfortunately, the world is facing growing water shortages, as reliable water availability is still taken for granted without considering its vast sustainability implications. However,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3111112, IEEE Access

access to water resources and fulfilling the need of water for the entire population is an open challenge. In developing countries, millions of people walk miles away in quest of safe drinking water. Moreover, the task of fetching water is often assigned to children and women, compromising education, health, and leisure, leading to improper growth and lack of personal development [2]. Therefore devising efficient and cost-effective solutions for accessing groundwater is crucial for ensuring food security and protection of living beings [3].

Drilling borewells to access water is a technology-intensive domain involving vast amounts of data requiring advanced data analytics techniques to extract knowledge discovery to bring improvements in drilling and water management and enable efficient decision-making. Over the past few years, the drilling industry has generated huge revenues, and currently, companies are investing billion dollars for research and development, rendering a time and resource-intensive process, as inefficient drilling operations could lead to potential resources loss.

The use of advanced data analytics aids in improving safety, cost-effectiveness, and quality of drilling operations. These problems can be analyzed and solved by applying different Machine Learning (ML) techniques [4]–[6]. The ML models have already penetrated critical decision-making processes such as predicting the timespan required to complete a drilling process at some specific location. It is also possible to predict how many days will be required to complete a drilling process in some specific location. This study mainly focuses on mining the groundwater bore drilling data to trace hidden patterns and relationships by clustering different regions based on distance, soil color, Korean layer, and water level. Moreover, for the discovery of distinct locations of boreholes, they are visualized in 2D and 3D contour and surface plot, i.e., point cloud [7]. The purpose of clustering the data is to get different patterns and regions suitable for the drilling. The dataset contains information for each bore location. Location is based on two coordinates, i.e., x and y. The bore location data is clustered into k distinct groups using different unsupervised techniques

As a data mining tool, clustering serves as a fundamental tool for combining similar data samples into k distinct clusters and gain valuable insights to data distribution for observing characteristics of individual clusters [8]. Previous research studies employed various clustering techniques that include Hierarchical Clustering [9], a distance-based approach that works by calculating the distance between the sample and the cluster. The idea of hierarchical clustering is to assign nearer objects to one cluster. Similarly, Balanced Iterative Reducing Clustering Hierarchies (BIRCH) [10] and Agglomerative clustering [11] are well-known algorithms that belong to this category. The partitioning method [12] is the second category of clustering family, which constructs k (where k < n) clusters and then evaluates the correctness of the method using some evaluation metrics, for instance, minimizing the mean square error. The third category of clustering is density-based clustering [13], which groups the dense areas based on the number of samples in the closed region. The last type of clustering is model-based clustering [14], which assumes that data is produced by a model and works by recovering the original model by using the data, thus making an attempt to finds the best fit of that model. The Gaussian Mixture Model (GMM) clustering is another widely used model-based clustering technique [15]. Seven clustering algorithms from all four types are used to cluster the different locations. Any particular clustering technique requires the number of clusters to be specified as an input, such as K-mean and Mini batch K-mean clustering [16]. For these clustering types, the first elbow curve is generated to get the optimal value of k. Affinity propagation and Mean shift are partitioning-based clusters, but the difference is that these methods don't require the number of clusters to be specified advance. The GMM is applied to distribute the data using the Gaussian distribution technique.

Likewise, in the case of unsupervised learning techniques, the evaluation criteria to measure clusters' effectiveness is based on the wellness of cluster formation considering factors such as intra-cluster distance and inter-cluster distance. For instance, the traditional K-mean clustering algorithm lacks the ability to cluster data when dealing with varied and dense clusters. Especially if data points are; scattered along with different densities and sizes, the mean can stumble towards the dense area. Hence due to dense data points, the radius of clusters turns larger; moreover, clusters will not be optimally separated. Our proposed method provides a solution to this problem by assigning a weight to each sample based on their distance from the mean value and calculate the weighted mean from that weighted feature space. The mean calculated from the transformed feature space shift towards the scattered points, resulting the inter and intra-cluster distance is minimized.

B. MOTIVATION

The quest for developing efficient and robust models in the field of exploration and developmental projects such as city construction and groundwater resource management with optimal drilling operations has ever continued. As discussed earlier, there is a dire need to optimize the efficiency of drilling operations for sustainable water resource management by minimizing the time and efforts required to improve the process of drilling for all stakeholders. As the process is resource-intensive involving costly equipment incurring multibillion-dollar budgets. Hence drilling optimization has a vital role in improving drilling performance that further helps in lowering the operational costs, drilling time and obtain superior performance in terms of high productivity and profitability. Furthermore, discovering land types in different areas and layers can significantly reduce risks such as stuck pipes, formation fracturing, and lost circulation. As different layers of the land possess various soil colors. These soil colors play an essential role if we have knowledge about the pattern of soil colors that lead us to the water. Moreover, the estimated number of days and the maximum depth should

also be known to plan the project and gain the water. By considering all these problems, we devised a solution based on the cluster formation with different structures of land, containing the information about the maximum depth, water level, and cost in terms of time.

C. CONTRIBUTION

The core contributions of this study are followed as:

- Soil color and land layer on different depths are analyzed, and average digging capacity per day is computed. The analysis aims to minimize the risk, such as stuck pipe, by selecting the area with soft soil and land layers
- Enhanced weighted K-mean is proposed in which the mean is computed from transformed weighted feature space
- Drilling time, productivity, and profitability factor are optimized by selecting the water level area on minimum depth.
- Dynamic feature sets are extracted, such as time interval of the drilling process, aggregated sum of borehole depths for each location, average digging capacity for each location based on soil color and land layer
- Comparative analysis is presented to show the effectiveness and significance of the proposed clustering algorithm based on extracted features groups.

The paper is structured as follows. Section II discusses the existing studies related to drilling process management; Section III presents the methodology of the proposed clustering approaches based on borehole depth and soil material. In section IV, we present the implementation environment, experimental and performance analysis results. Section V concludes the paper with possible future direction.

II. RELATED WORK

The scarcity of groundwater resources is becoming a global challenge. Over the past few years, the surge for groundwater exploration has risen beyond limits. The growing water demand has led to more water extraction through the drilling process [17]. Recently drilling has seen significant developments to cater to the vast water needs worldwide. The drilling industry consumes massive budgets because of its multidisciplinary nature, a requirement of the skilled task force, and dynamic real-time operations [18]. Gaining water has so many application areas; some include exploration of a non-renewable resource, construction sector needs drilling for underground projects, mining sector and engineering sector and geotechnical research projects also require drilling operations for their projects. As the drilling process for groundwater extraction requires the latest machines and tools that are costly and hardly affordable for developing countries. Over the past few years, the drilling industry has evolved into a multibillion-dollar industry. Therefore, the water gaining process must be optimized and given thorough attention to avoid any resource loss [19]. An uncontrolled and inefficient process could pose a potential threat to a country's economy and sustainability. Hence, groundwater exploration through the bore log process must be time and cost-effective. Additionally, the process involves making complex and optimal decisions; inappropriate decisions can significantly impact the performance and cost [20].

Like all other branches of science, drilling, hydrology, and geosciences are also undergoing breakthrough transformations based on research, technological advancements, and big data analytics [21]. Big data analytics have revolutionized these research domains through advanced analytical techniques for leveraging a massive amount of heterogenous drilling data [22]. The explosive rise in drilling groundwater resources and advances in drilling tools generate massive data, and managing such data is a significant concern of drilling companies. Big data analytics is a powerful tool for managing and processing vast amounts of drilling data to reveal underlying hidden patterns and equations related to sophisticated drilling groundwater processes [23].

Big data analytics and ML methods are highly preferred and adopted by scientists for geoscience datasets. For instance, the authors in [24] employed ML techniques for the scientific ocean drilling dataset. ML methods are categorized as supervised and unsupervised. Unsupervised learning has no response variable; it only attempts to find the hidden patterns in input data. In contrast, supervised learning has target variables and labeled inputs. Data exploration is the process of finding patterns and identifying trends in data regardless of prior knowledge [25]. For discovering a hidden pattern in drilling, data clustering is a suitable technique, as it can identify the density and sparsity of particular regions in a dataset having their attributes—clustering group similar objects into one group by calculating the distances between objects. In [26], the authors presented a maximum likelihood-based approach for clustering to improve drilling performance. The proposed approach generated patterns for recommending optimal drilling parameters. Another study [27] presented a cluster-based analysis for classifying groundwater wells based on water quality. In another study [28], the authors presented a correlation analysis between drilling parameters and geological parameters of rock and soil by considering mechanics and energy factors.

Nowadays, statistical analysis-based methods are also being extensively applied to the drilling domain due to their computationally inexpensive nature and non-requirement of physical application scope. In [29], the authors proposed an automatic drilling hazard detection method based on statistical analysis. Groundwater drilling data comprise continuous time-series data. The analysis of drilling time-series data is done as whole rather than individual parts. As drilling and hydrogeological time series data possess a dynamic behavior and their physical and chemical properties change over time [30]. Fuzzy clustering approaches are well-known clustering techniques applied to hydrogeological drilling data due to their ability to provide extra information related to membership degrees and variation detection in various hydrogeological parameters [31]. For instance, data for classification

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3111112, IEEE Access

applied a fuzzy logic-based clustering technique to cluster data samples [32]. Clustering is an unsupervised collection of data exploration methods employed to group together naturally occurring similar objects [33]. The authors proposed a method to determine the relationship between various drilling parameters by applying K-means clustering. The proposed work aims to optimize the drilling parameters based on conditions resulting in high penetration rates [8]. Experimental findings suggest direct and inverse relations among various drilling parameters. The authors employed partitional cluster analysis to identify the relationship between hydraulic connectivity, lithology, and geotechnical attributes [34].

Clustering approaches are extensively applied to multivariate hydrogeological and drilling time-series data to discover knowledge, and hidden patterns among such datasets [35]. Cluster analysis comprises two basic approaches that are variable clustering and partitional clustering. C-mean and fuzzy c-mean are the most widely used partition clustering algorithms for drilling and hydrogeochemical data [36]. Varying clustering techniques define multivariate relationships among data, while partition-based clustering techniques assign samples to specific groups.

Classification algorithms are often applied to the exploration of underground non-renewable resources. Firstly, these methods train the classifier using drilling data acquired from the preliminary and detailed drilling process [37]. Afterward, the trained classifier is evaluated by classifying test data by assigning them to various classes. Thus, classification methods are efficient at providing relevant results comparative to cluster-based analysis. However, training data requirements for classification restrict such algorithms for drilling and hydrogeological datasets. Additionally, the availability of drilling data is possible because of the real-time drilling data process and is often scarce. Hence, the classification model cannot provide precise and reliable results [38]. Therefore to overcome the problems in existing classification techniques, unsupervised and semi-supervised clustering techniques are an efficient choice.

Although many existing techniques are applied to enhance the efficiency and planning of borehole resources. However, due to variations in hydrological patterns, it is still an open research area to investigate different characteristics of the borehole process to facilitate drilling management. The proposed model utilizes different clustering techniques to divide the regions based on different hydrological characteristics. The ultimate goal of the study is to help the drilling industry figure out the region's situation before starting the drilling process. Furthermore, underground surveillance management can investigate the water level, state of the soil material, and the whole process's cost. The proposed L2 Weighted K-mean is applied and compared with traditional machine learning algorithms. For future the ensemble clustering techniques [39], [40] can also be applied on different feature sets to extract hidden patterns from the data.

TABLE 1: Description of the dataset

Attribute	Description
X location	x coordinate of the drilling point
Y location	y coordinate of the drilling point
Starting depth on a specific day	Shows the start of the drilling depth
Ending depth on a specific day	Shows the end of the drilling depth
Total depth	Derived attribute from ending depth Shows the total depth of each single drilling point
Ground water level different locations	The value shows the water level on
Number of days single bore.	Shows total days spent on each
Korean and layers discrete values	The layer of the land having nine
Soil color	The soil color on different depth. Having ten different soil colors.

III. PROPOSED CLUSTERING APPROACHES

This section presents a methodology of the proposed clustering approaches based on drilling depth and soil materials. Figure 1 shows the basic flow of the proposed approach. The methodology of the proposed study consists of the following steps; acquisition of drilling dataset, preprocessing, features extraction and grouping, ML-based clustering approaches, performance evaluation, and visualization of resulting data in a 3D format for better understanding.

A. BORE-LOG DATA

The proposed study employed a real borehole drilling dataset acquired from Jeju National University, Republic of Korea. A visual representation of the dataset containing attributes related to drilling is shown in Figure 1. Each feature contains different characteristics of drilling points, including location, depth, and soil color. As each attribute has different characteristics of drilling point so, the attributes are grouped into six different feature groups. Each feature group contains different drilling information like level of water in different areas, soil color on different depths, Korean layer on different locations. All the extracted feature groups from MySQL workbench are shown in Figure 2. All the attributes which are considered in the experiments are listed in Table 1.

B. PREPROCESSING OF BOREHOLE-LOG DATA

Preprocessing is one of the essential steps of the experiment leading to a cleaner, meaningful and manageable datasets. Hence data transformation is vital for meeting the requirements of ML models. Various methods are used to remove unwanted data and fill missing values to increase the reliability of the dataset. Before passing the data to the clustering model, it is required to process data samples in order to transform raw data samples into a reliable format. As we have categorical features, the string values are encoded to numeric by using the ordinal encoder [41]. In an ordinal encoder, a unique number is assigned to each unique category; for

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3111112, IEEE Access

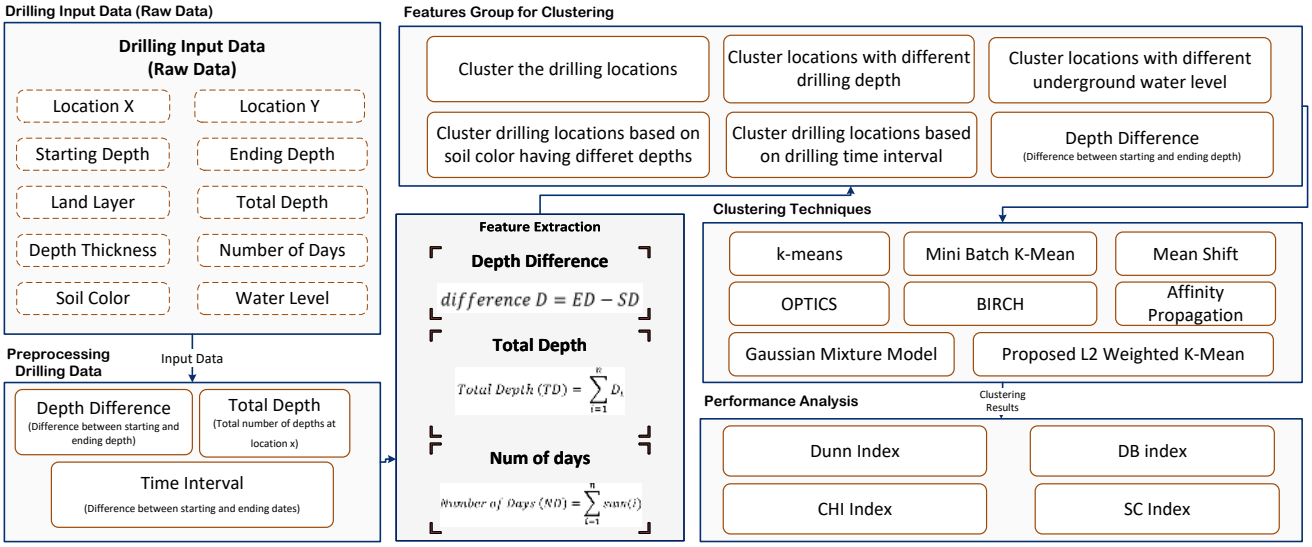


FIGURE 1: Basic flow of the proposed clustering approaches

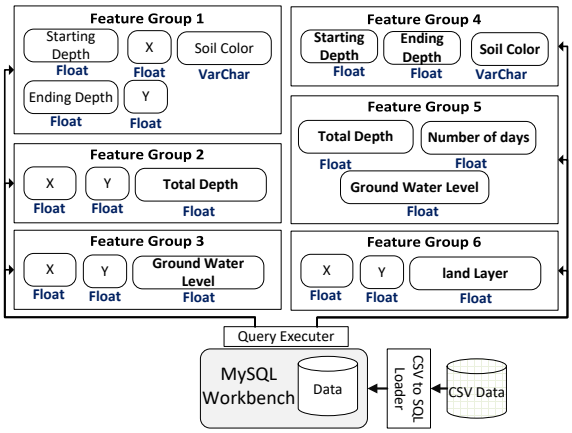


FIGURE 2: Feature groups retrieved from MySQL Workbench

instance, we have ten soil colors and nine land layer names as a string value, so the ordinal encoder assigns values from 0-10 and 0-9, respectively. After encoding the data, some features are dropped. The dropped features have no pattern and possess a unique effect for each sample of the data, e.g., drilling resonance and borehole code, which is unique for each sample. Hence relevant features are selected, and other features are dropped.

After the feature selection, we dealt with missing values as they can drastically affect the performance of ML algorithms. Dealing with null values before applying any algorithm is essential because they generate an error during the calculation. One way of dealing with missing values is to remove the samples with null values, but generally, its a non-preferred choice. Missing values can be tackled using various techniques depending upon the nature of the problem. In this study, a k-Nearest Neighbor (KNN) imputer is utilized as a standard technique to fill the missing values [42]. This method uses the KNN technique to replace the missing values with the calculated value. The KNN based imputer works by calculating value is based on the mean of its selected neighbors. In our case, the minimum number of instances for each borehole is 3, and the maximum is 9, so the average number (6) is used to fill the missing value. The Euclidean distance is the default distance metric to impute the missing values. The flow of feature selection, extraction, and preprocessing is shown in Figure 1.

Once the data is preprocessed, all the features are selected, and the data is passed to the features extraction and grouping phase.

C. FEATURES EXTRACTION AND GROUPING

The data contains multiple features, including starting depth of the bore, ending depth of the bore, and location, i.e., x and y coordinates. To extract the new features from the existing one, data is analyzed in geological terms. The first feature which is computed from the data is the depth difference. The depth difference is computed as the difference between ending and starting drilling depths. The starting depth (SD) shows the depth at which drilling is started that day; similarly, the ending depth (ED) is the end of the depth in meters on the same day, as shown in Figure 3. The depth difference is computed as shown in equation 1.

$$difference(D) = ED - SD \tag{1}$$

where ED is ending depth and SD is the starting depth.

The total depth is the sum of the difference between ending and starting depths for drilling location i. The total depth is

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3111112, IEEE Access

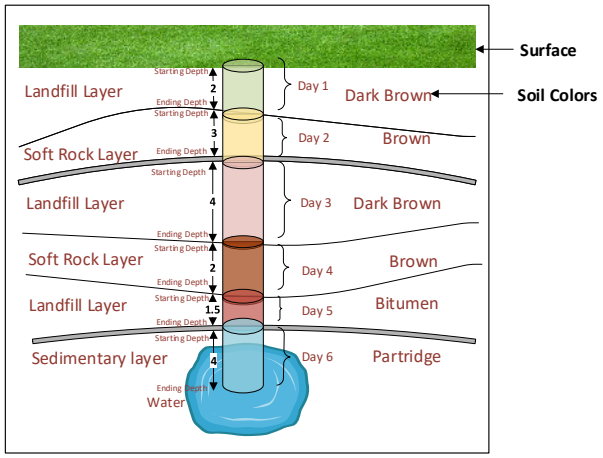


FIGURE 3: One sample of Drilling point with number of days, soil color and land layer

the depth of the drilling location i where the water is gained. The total depth is calculated as shown in equation 2.

$$totaldepth(TD) = \sum_{i=0}^{n} D_i \tag{2}$$

where D is $depth\ difference$ calculated from equation 1.

Another feature is computed as called number of days spent by counting the instance of one borehole. Different depth range is achieved in different days as shown in Figure 3.

D. DESCRIPTIVE ANALYSIS OF BORE-LOG DATA

Data analysis is the process of analyzing hidden patterns and characteristics of the dataset. This study applies data analysis techniques to discover underlying patterns related to soil color and land layer. We discovered different soil colors acquired during the drilling process: the data analysis yielded ten different soil colors and eight significant land layers in the given data. As time factor is critical to drilling operations, it is essential to know how many days will be spent on each soil color and land layer. Figure 4 shows the relationship between the number of days spent and the total depth achieved on each soil color. It is evident from the figure that the soil with brown color has the highest average digging rate per day and consumers more days comparative to other soil colors. The Average digging capacity per day on each soil color is also mentioned. The average depth reflects the hardness level of the soil color. The hardness of the soil is an essential factor to know to reduce the risk of stuck pipes. The dark brown soil color is spongy and soft compared to the rest of all because the digging capacity incurred by the dark brown soil is 4.34 meters per day. Moreover, partridge and gray soils are more challenging than all others because the digging capacity is 2 meters in both cases. The soil color with less value of total depth also illustrates the thickness of the layer of that

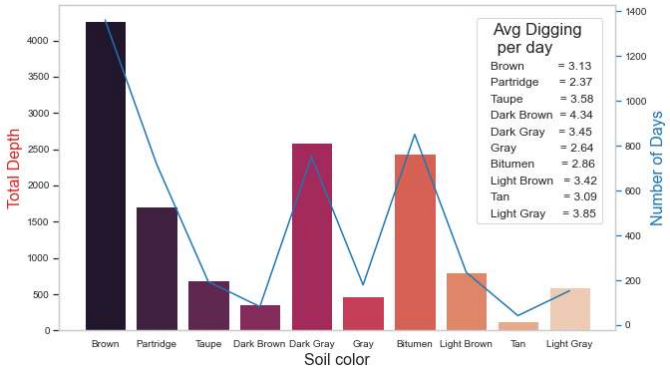


FIGURE 4: Analysis of soil color based on number of days, maximum and average depth achieved per day

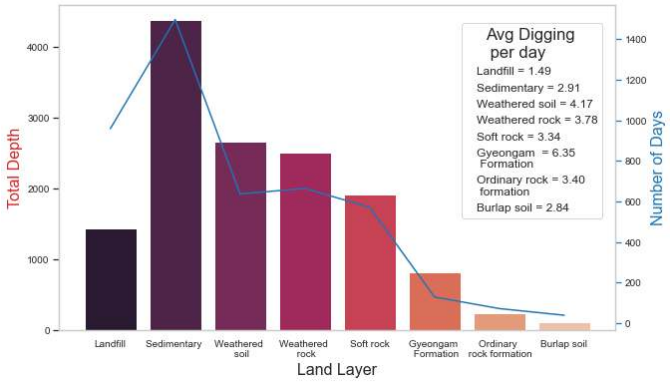


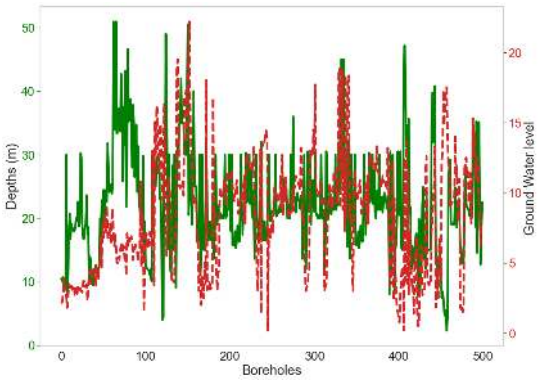
FIGURE 5: Analysis of Korean land layer based on number of days, maximum and avg depth achieved per day

soil color. The soil with tan color is rare in Korea because the total amount of soil found with tan color possesses 165-meter depth. Hence brown color stands first in achieving the highest depth, and this indicates that the brown color is the most expected occurring soil color in the drilling process.

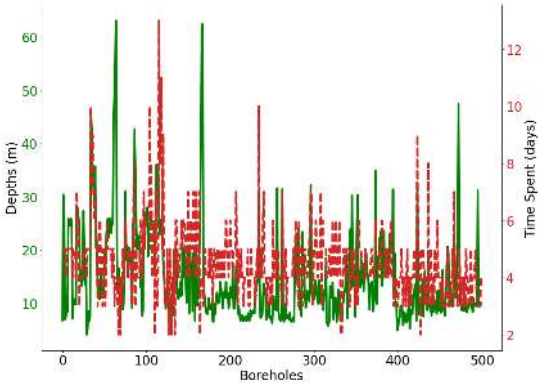
Likewise, the land layer also holds attributes such as hardness and depth range per day, similar to soil color. Figure 5 shows that the sedimentary layer is most common in Korea; moreover, the layer's digging capacity is 2 meters, which shows that the layer is too hard compared to others. On the flip side, the burlap layer is rare, but the hardness level is the same as the sedimentary layer. The hardness of the layer can be seen in the same figure, which shows one-day digging capacity of the particular land layer. The digging capacity on the Gyeongam formation layer is double the ordinary rock layer. The softness level of the landfill layer, weathered rock layer, soft rock layer, and ordinary rock layer is the same because the digging capacity is 3 meters per day for all of these layers.

Each unique drilling point has different characteristics in terms of days, depth, and water level. Figure 6a shows the relationship between groundwater level and total depth of different boreholes. Each drilling point has its own different depth, which shows the water level from the ground and the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3111112, IEEE Access



(a)



(b)

FIGURE 6: Relationship between Total depth of borehole with water level and number of days spent

right y-axis of the figure shows the level of the water on that depth. In most of the locations, water is achieved on 10 meters, while the water level is too low in some areas. The number of days with the total depth of different drilling points is given in Figure 6b. The illustration shows that the maximum number of drilling points are finished within 4-6 days. The maximum number of days spent on any drilling point is 14.

The acquisition of groundwater involves digging through a sequence of layers and soil colors that ultimately leads us to the water. To ascertain the sequence, samples of soil color and the Korean layer are grouped. To aim this objective, we grouped and visualized the samples based on starting and ending depth. The resulting sequence is visualized and illustrated in Figures 7.

The samples of the Korean layer are grouped, and their mean is visualized based on the frequency of the samples within one group. Starting depth is on the x-axis, while the y-axis shows the ending depth. At first glance, it can be clearly seen from Figure 7 that the Landfill layer is the most commonly occurring layer encountered by the drilling industry at the start of the digging process. The sedimentary layer is followed by the soft rock layer, while the soft layer is less common because the frequency measure of that layer is

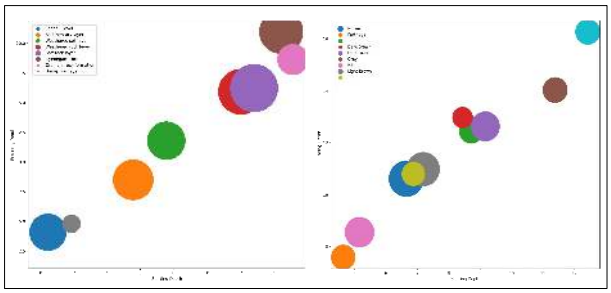


FIGURE 7: Hidden patterns of soil color and land layers

lower than all others. When Gyeongam formation or ordinary rock formation layers appear in the drilling process, there is a high probability that the water could be approached after these layers. Now, by taking into account the pattern of soil color, the brown color is a sign of water discovery in the near premises. Usually, water can be found after that. Similarly, each Korean layer can relate the soil color to get the color of the Korean layer, e.g., landfill layer is in the color of partridge or bitumen because these are appearing on the same depth. The overlapping of layers and soils colors are showing the mixed pattern on the same depth.

The visualization of the extracted features on a 3D surface plot is done to analyze the patterns. The underground water table is at a different depth from the ground surface. The pattern visualized in Figure 8 shows the water table. The figure also presents the information from where the water is started. In some areas, the water level is high, while in others, it is low.

The attributes are displayed in contour and surface plots to get the whole surface's view. The given data is visualized in 3D space by selecting some features and target attributes. The entire surface of the given data is shown in Figure 9, along with the maximum depth of each drilling point. The surface shows that some areas have water at maximum depth while the water is on minimum depth in some areas. The blue area on the map shows minimum depth or 0 depth (no drilling point). The red peaks of the surface indicate the maximum depth of any drilling point. The map is plotted on x and y coordinates of the given data.

Moreover, a 2D surface plot is also visualized to partition the areas based on different characteristics. Figure 10a shows the soil color on different regions. We have ten distinct soil colors in the dataset, listed on the right bar of the figure. Multiple soil colors are found at different depths of drilling points. The underground quantity of soil color is also different in different areas. By examining the quantity of soil in different places, Figure 10a is plotted. Similarly, each drilling point's depth is varied at different locations, illustrated in Figure 10b. For the sake of clarity, the regions are also divided by various depths of drilling points. The right color bar presents the depth range with a different shade.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3111112, IEEE Access

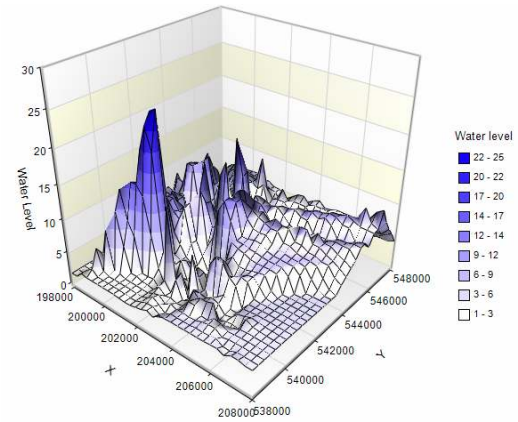


FIGURE 8: Analysis of water level on different locations

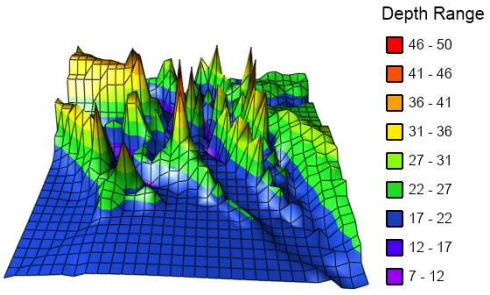


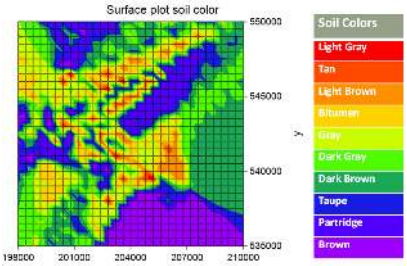
FIGURE 9: Point cloud visualization of total depth on different locations

Furthermore, the map is also divided based on these depths. The blue area of the map shows the smallest depth or water level at minimum depth, and the reddish area determines the highest depth of drilling points.

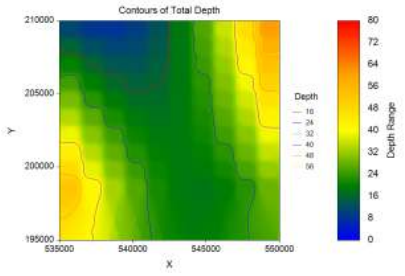
E. CLUSTERING APPROACHES

Once features are obtained and grouped, clustering techniques cluster the input data samples into distinct groups based on hidden characteristics and patterns. All six feature combinations are given as input to the clustering model. In this study, several clustering algorithms are implemented, and results are compared with the proposed weighted K-Mean clustering algorithm to get desired results to improve the overall drilling process. First, K-mean and mini-batch K-mean are applied to cluster drilling data using extracted feature groups. These partitioning algorithms divide the data based on the value of k, required in the first step.

K-mean clustering is the more commonly used clustering algorithm in many unsupervised learning problems. The K-means clustering algorithm's main objective is to groups the data samples into k distinct groups and discover underlying patterns. K-means looks for a constant k to cluster the data and achieve the objective. Therefore, it is needed to determine the total number of clusters for the prepared dataset. In this study, an elbow curve method is employed as a standard



(a) 2D contour plot shows the soil color on different areas



(b) Maximum depth range on different areas

FIGURE 10: Contour plots show the soil color and maximum depth on different locations

to determine the optimal value of k. The elbow curve method is the fundamental step for unsupervised learning to determine the optimal value of k. After determining the optimal value of k, the value is passed to the K-mean model. Initially, K-means select the initial centroids randomly and group the samples based on Euclidian distance. The following equation 3 is used to calculate the distance between each data sample x_i and each cluster centroid c_i [43].

$$distance(x_i, c_i) = \sqrt{\sum_{i=1}^{n}(x_i - c_i)^2} \tag{3}$$

Figure 11 shows the elbow curve analysis based on extracted features groups using K-means clustering. It can be observed that optimal k is 2 or 3 for each feature group, which indicates that the given data samples are grouped into two or three distinct groups. This elbow curve is helpful in terms of partitioning clustering techniques because these clustering techniques require the number of partitions in advance.

Another clustering algorithm that is similar to K-mean is mini-batch K-mean. The difference is that it first fragments the data into multiple batches. The main idea of the algorithms is to use small random batches of data of fixed size so that they can be stored in memory easily. In each iteration of the mini-batch K-mean, a new random sample from the dataset is obtained and used to update the cluster and repeated the steps until convergence. The mini-batch processing is faster than traditional K-mean clustering because of small batches. The primary objective of applying mini-batch K-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3111112, IEEE Access

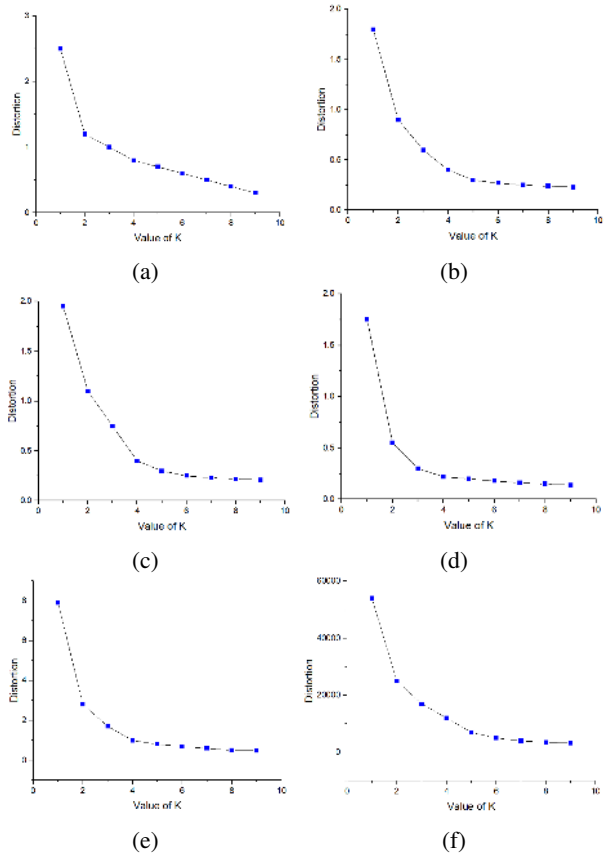


FIGURE 11: Elbow curves of all six feature groups

mean is to cluster the data samples faster than the traditional K-mean clustering algorithm.

The density-based clustering algorithms, i.e., Mean shift (MS) and AP accept the processed FGs and cluster the data based on dense area. These types of algorithms do not require the number of clusters in advance. MS algorithm first sets the window randomly on the data and then extends the window based on the distance of the input samples. On the other hand, AP constructs a similarity matrix by deciding the exemplar and clusters the dense areas of the surface. BIRCH and Agglomerative Clustering techniques generate the hierarchical flow in Feature Tree (FT) and Dengrogram, respectively. From the model-based clustering, GMM is potent and extensively applied in various tasks requiring data clustering.

$$ln\ p(X) = \sum_{n=1}^{N} ln \sum_{k=1}^{K} \pi_k N(x_n|\mu_k, \sum k) \qquad (4)$$

Gaussian distribution, a model-based clustering technique, is used to distinguish between the samples in the GM model. GM first compose the model and apply gaussian distribution to similar group samples. By minimizing the likelihood (4) GMM finds the optimal value clusters [44].

OPTICS is a density-based hierarchical clustering technique that identifies absolute shaped clusters and reduces

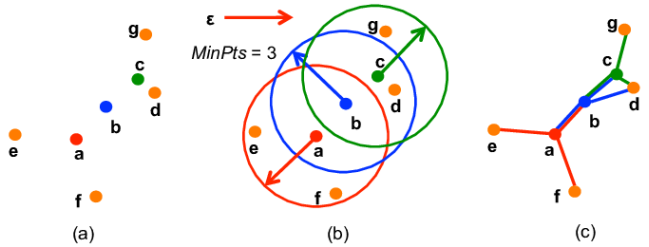


FIGURE 12: Hierarchical density based selection of points by Ordering points to identify the clustering structure (OPTICS)

noise using flexible reachability measure thresholds [45]. The OPTICS algorithm is considered challenging because it exhibits an entirely regular data access plan. Optics initially select arbitrary points from the dataset and then identify all reachable density points concerning EPSI value and minimum neighbor points, as shown in Figure 12 . OPTICS accepts these two attributes in advance.

Algorithm 1 is used to presents the basic flow of the clustering approach. The data passed to the algorithm is fetched from the MySQL database, and the output of the algorithm is optimal clusters based on different characteristics of the land area. First of all, data is divided into different feature groups that represent hidden patterns and characteristics. Then, by getting the optimal value of k, an elbow curve-based heuristic method is utilized to determine optimal k. Data samples along with optimal k are then passed to the clustering algorithm. In data preprocessing, the KNN imputer is used to fill missing values, and the ordinal encoder is used to encode values from string to integer. Next, by initializing the clustering algorithms, each technique is used to cluster the data. Lastly, different evaluation measures are utilized to evaluate the implemented clustering techniques, such as Dunn Index, DB Index, SC, and CHI, by getting the clusters from the clustering technique.

F. PROPOSED L2 BASED WEIGHTED K-MEAN CLUSTERING MODEL

K-mean clustering algorithm clusters data by using the mean of the cluster and then compute the distance from the mean to any sample. The smallest distance from the mean shows that the sample belongs to that particular cluster. However, when we have dense and scattered points in different clusters, the mean tends to shift towards dense areas, due to which some samples can be incorrectly clustered. Likewise, if clusters are far apart from each other, the mean strategy can work, but in the case of mixed clusters, the mean will not be an effective solution. Hence the optimal number of clusters is chosen based on the elbow curve and passed to the proposed model for further processing.

Different reservoir-based and hash-based sampling algorithms are proposed to overcome biased density sampling on large dataset [46]. The reservoir sampling algorithm is unbiased and random. The algorithm is used in many

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3111112, IEEE Access

Algorithm 1: Clustering the Data

Data: Training set S = (x_1,x_2,x_3,...x_n) drilling Data
Result: Optimal Number of clusters based on different characteristics of land
initialization;
data $\leftarrow (RequestSQL)$;
$FeatureGroups \leftarrow SplitFeatures(data)$
$k \leftarrow ElbowCurve$
for *each FeatureGroup* **do**
  $FeatureGroup \leftarrow KNNImputer(FeatureGroup)$
  if *AlphaNumericFeatures* **then**
    Ordinal Encode Features;
    Merge with existing features;
  $ClusteringAlgorithms \leftarrow InitilizeParameters()$
  for *each* ***ClusteringAlgorithms*** **do**
    $starttime \leftarrow currenttime()$;
    $Clusters \leftarrow ClusteringAlgorithm(FeatureGroup)$;
    $endtime \leftarrow currenttime()$;
    Evaluate;
    $DunnIndex(U) \leftarrow \min_{1<i<c}\left\{\min_{1<j<c}\left\{\frac{\delta(X_i,X_j)}{\min_{1<k<c}\{\triangle X_k\}}\right\}\right\}$
    $DBIndex(U) \leftarrow \frac{1}{k}\sum_{i=1}^{k}\max_{1<i<c}\left\{\frac{\Delta(X_i)+\Delta(X_j)}{\Delta(X_i,X_j)}\right\}$
    $SC \leftarrow \frac{b-a}{max(a,b)}$
    $CHI \leftarrow \frac{SS_B}{SS_w}\times\frac{N-k}{k-1}$
    $Time \leftarrow endtime - starttime$;

areas to assign weights to scan the dataset and assign weights—similarly, the hash-based [47]–[49] . The genetic K-mean (GWKMA) [50] is also proposed and uses a Genetic Algorithm (GA) to assign weights to the samples. In hash-based and reservoir-based weighted k-mean clustering methods, the weight function is used to scan the data at once and compute the weighted mean of each cluster. The scan is done after completing clusters using traditional k-mean, while in the proposed weighted K-mean clustering algorithm, weights are assigned to each sample of the data. The weights are based on the distance of that sample from the mean value. The first step of the proposed method is to compute the mean by using the following equation.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{5}$$

Where $\bar{x}$ represents the mean of the cluster, n is the total samples in that particular cluster, and x_i is the sample of that cluster.

After calculating the mean value, the $\bar{R}$ radius of the cluster is computed. The value of radius is the maximum distance of any sample from the mean value i.e.

$$\bar{R} = max_i \|x_i - \bar{x}\|_2^2 \tag{6}$$

the radius of the cluster shows the area covered by that particular cluster. next weight for each sample within the cluster is computed by using

$$weight\ x_i = \|\bar{x} - x_i\|_2^2 \tag{7}$$

where $\bar{x}$ is the mean value, and x_i is the sample for which the weight is being computed.

The weights are assigned to each sample to transform the original feature space into a weighted feature space. The weight to each sample is assigned by using the following formula

$$\bar{x}^{(w)} = \frac{\sum_{i=1}^{n} x_i.weight\ x_i}{\sum_{i=1}^{n} weight\ x_i} \tag{8}$$

After transforming the samples from the original to weighted feature space, the mean is again computed by using 5. The weighted mean computed from the weighted feature space reduces the area covered by the cluster and creates the optimal cluster. The distance between clusters is evaluated using a radius of the cluster, while the distance among the cluster is evaluated using DUNN and DB index. The mathematical model of the proposed weighted K-mean algorithm is shown in Figure 13.

Algorithm 2 shows the proposed model based on two steps. The first step is to apply K-mean to get the mean value of the cluster and then transform the data into a weighted feature space to calculate the weighted mean.

IV. EXPERIMENTAL ENVIRONMENT AND CLUSTERING RESULTS

This section presents the experimental environment, clustering results, and performance analysis.

A. EXPERIMENTAL ENVIRONMENT

Experiments are conducted on a Windows PC with 12GB RAM. A front end (Desktop application) is developed using Java, and the clustering techniques are applied in python. Well-know python libraries, including NumPy, SkLearn, and Scipy, are used for clustering experiments. In addition, NCSS is used for the visualization of data in PC. The required software and hardware components are listed in Table 2.

First, the preprocessed data is loaded from CSV to MySQL Workbench using the data loader SQL command. Next, the chunks of data are picked from the table by using SQL query. The query is directly executed from the simulation environment, i.e., python (Anaconda). As the data is divided into six feature groups, each has an SQL procedure that returns related attributes. By calling SQL procedure from the simulation environment, data is available in the form of a data frame in a simulation environment. Next, all the feature

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3111112, IEEE Access

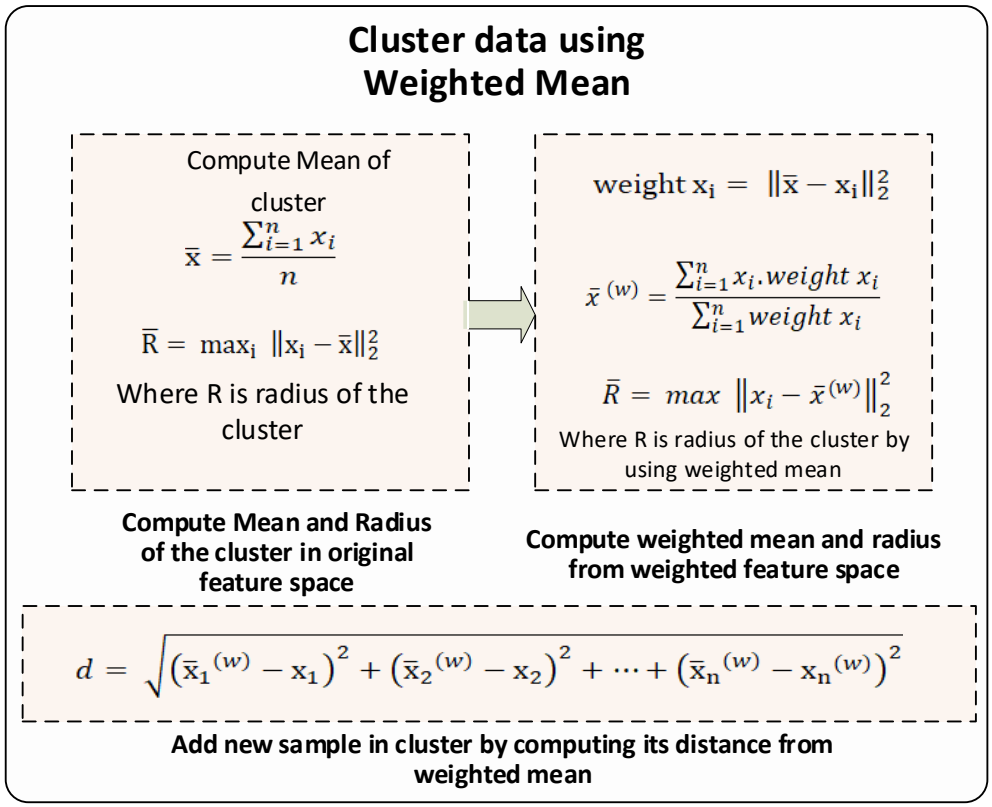


FIGURE 13: Mathematical model of the proposed weighted K-mean clustering algorithm

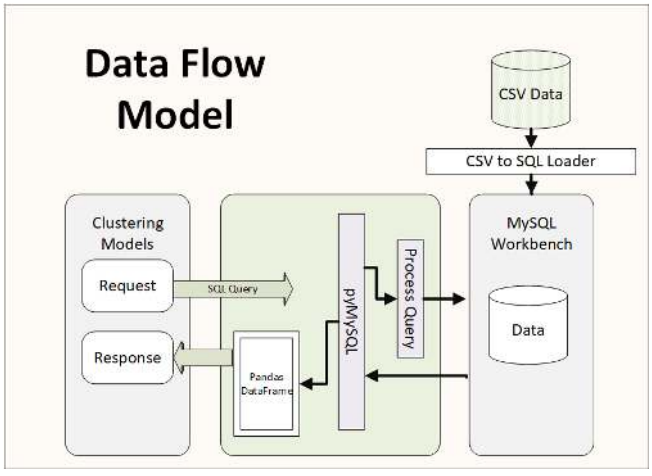


FIGURE 14: Flow of the data from MySQL to Clustering algorithm

groups are passed to clustering models as shown in Figure 14.

B. CLUSTERING RESULTS AND ANALYSIS

We applied clustering evaluation metrics to evaluate the performance of clustering algorithms in terms of the division of areas based on different characteristics. These clustering metrics are suitable when the ground truth label is not given.

Dunn Index is a metric to evaluate clustering algorithms. Like all other cluster evaluation metrics, identifying the well-separated clusters and minor variance between members of clusters is the main strategy of the Dunn index. Moreover, it evaluates the intra-cluster variations and checks if the mean of clusters is sufficiently far apart. The higher value of the Dunn index shows the better separation of the clusters. The optimal number of k can maximize the value of the Dunn index. The Dunn index is not high because multiple soil colors and the Korean layer lie in the same location. The Dunn index with c number of clusters is defined as,

$$DunnIndex(U) \leftarrow \min_{1<i<c} \left\{ \min_{1<j<c} \left\{ \frac{\delta(X_i, X_j)}{\min_{1<k<c}\{\triangle X_k\}} \right\} \right\} \quad (9)$$

where,
$\delta(X_i, X_j)$) is inter-cluster distance i.e., distance between cluster X_i and X_j. $\triangle X_k$) is the intra-cluster distance of cluster X_k i.e., distance within the cluster X_k.

Algorithm 2: Proposed 2 Step Weighted K-Mean Clustering Algorithm.

Data: Training set S = $((x_1),(x_2),,...(x_n))$, k as number of cluster
Result: Optimal Clusters and the Radius of each cluster
initialization;
Initialize cluster centroids $\leftarrow \mu_1,\mu_2...\mu_k \in \mathbb{R}^n$;
Folds $\leftarrow SplitDatatoFolds(S)$
for $x_j \leftarrow S$ **do**
  Step 1:
  Compute Mean Value
  $\mu_k = \frac{\sum_{i=1}^{n} x_i}{n}$
  For new sample x_j, compute distance from centroid of each cluster
  $d = \sqrt{\sum_{j,k=1}^{n}(\mu_k - x_j)^2}$
  Assign cluster based on minimum distance
  Step 2:
  Compute Radius
  $\overline{R} = max_i \parallel x_i - \mu_k \parallel_2^2$
  Compute weight for each sample
  $weightx_i = \parallel \overline{x} - x_i \parallel_2^2$
  Weighted Feature Space
  $\mu_k^{(w)} = \frac{\sum_{i=1}^{n} x_i.weightx_i}{\sum_{i=1}^{n}.weightx_i}$

TABLE 2: Experimental setup and required components

Sr#	Component	Description
1	Hardware	PC
2	Operating System	Window 10
3	Memory	8GB
4	Server	SQL Server
5	Libraries	Pandas, Mysql, SKLearn,
6	Front End	Java
7	Storage	MySQL, MS Excel
8	Core Programming Language	Python, Java, SQL
9	IDE	Anaconda Navigator (3) Jupyter Lab

Davies–Bouldin index (DBI) is another clustering metric used to evaluate the clusters based on internal evaluation schema, where the validation of how well the clustering has been done is made using quantities and features inherent to the dataset. The DB index with k number of clusters is defined as,

$$DBIndex(U) \leftarrow \frac{1}{k}\sum_{i=1}^{k} \max_{1<i<c} \left\{ \frac{\Delta(X_i) + \Delta(X_j)}{\Delta(X_i, X_j)} \right\} \quad (10)$$

where,
$\Delta(X_i)$ and $\Delta(X_j)$ are intra-cluster distance of X_i and X_j respectively.$\Delta(X_i, X_j)$ is inter-cluster distance i.e., distance between cluster X_i and X_j.

Silhouette coefficient (SC) is another evaluation metric used here to evaluate the performance of clusters. The higher value of SC defines better cluster separation. SC for a sample is

$$SC \leftarrow \frac{b-a}{max(a,b)} \quad (11)$$

where b is the distance between a sample and the nearest cluster that the sample is not a part of it. The a is the mean intra-cluster distance
Calinski–Harabaz index (CHI) is the metric to evaluate the clustering algorithm. CHI can be computed by

$$CHI \leftarrow \frac{SS_B}{SS_w} \times \frac{N-k}{k-1} \quad (12)$$

Where k represents the number of clusters generated by the algorithm, and N is the total number of observations, i.e., data points within the cluster. SS_B represents the overall intra-cluster variance, and SS_W represents inter-cluster variance.

The clustering techniques are applied to six different feature groups, and the area is divided based on different attributes of the land. The first feature group kept the location information along with soil color and starting and ending depth. This feature group reflects the information related to soil colors on different depths at distinct locations. These clusters can be used to pick the area with soft soil. The value of the Dunn index is based on the distance between clusters, i.e., intra-cluster distance. Due to the same color on the same depths in different regions, these clusters may overlap each other for some feature value, and the value of the Dunn index is low. Figure 15 shows the example data divided into different clusters. The illustration makes the fact clear that weighted mean of sample shift towards the scattered points of the cluster. It is also evident that the clusters defined by L2 weighted K-mean are more well defined in terms of inter-cluster and intra-cluster distance.

The different number of clusters are produced by algorithms applied on different feature combinations as shown in Figure 16. The type of algorithms that accept the number of clusters as input, including K-mean, Mini-batch K-mean, and GMM, computes the value of k based on the elbow curve. An elbow curve is generated to get the number of optimal clusters, and the value of k is then passed to the partitioning algorithms. The clusters generated by affinity propagation and OPTICS are too high because the dense area influences the cluster separation. The division of groups based on distance is too high in all feature groups. These clusters provide little information of the area based on the characteristics of the land. For example, the soil color and land layer have different hardness levels in different regions. The soil, water level, total depth, and land layer analysis show the suitable areas for drilling. The drilling industry should not select the cluster with hard soil or layer, or the related resources should be managed before starting the drilling process. The hard surface can take time to drill, and the risk of pipe stuck is also involved. Because of the hard surface, the drilling fluid can be broken, and the labor cost is also high because of the slow drilling speed.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3111112, IEEE Access

TABLE 3: Comparison of proposed clustering approaches using extracted features groups

		Indexes			
Features Groups	**Clustering Techniques**	**Dunn Index**	**DB Index**	**SC Index**	**CHI Index**
Feature Group 1	OPTICS	0.034	1.235	0.473	27
	Mini Batch K-mean	0.051	0.967	0.729	3325
	K-mean	0.059	1.019	0.625	3273
	BIRCH	0.046	1.522	0.763	1316
	Affinity Propagation	0.167	1.063	0.293	1874
	Gaussian Mixture Model	0.046	1.524	0.163	1312
	Proposed L2-W K-mean	**0.267**	**0.563**	**0.948**	**3645**
Feature Group 2	Mean Shift	0.215	0.821	0.849	946
	OPTICS	0.079	**0.741**	-0.223	27
	Mini Batch K-mean	0.122	0.956	0.792	879
	K-mean	0.193	0.881	0.637	973
	BIRCH	0.171	0.947	0.817	938
	Affinity Propagation	0.215	0.861	0.398	968
	Gaussian Mixture Model	0.036	1.159	0.344	687
	Proposed L2-W K-mean	**0.318**	0.789	**0.879**	**1145**
Feature Group 3	Mean Shift	0.068	0.57	0.594	1101
	OPTICS	0.017	1.788	0.124	118
	Mini Batch K-mean	0.157	0.596	0.828	1119
	K-mean	0.173	0.464	0.759	1474
	BIRCH	0.263	0.568	0.864	1109
	Affinity Propagation	0.211	0.486	0.605	2225
	Gaussian Mixture Model	0.023	0.897	0.399	3046
	Proposed L2-W K-mean	**0.293**	**0.318**	**0.871**	**3179**
Feature Group 4	Mean Shift	0.049	0.554	0.683	6589
	OPTICS	0.006	2.406	-0.305	32
	Mini Batch K-mean	0.036	0.762	**0.985**	1636
	K-mean	0.019	0.701	0.816	6331
	BIRCH	0.013	1.325	0.565	4555
	Affinity Propagation	0.02	0.777	0.395	4595
	Gaussian Mixture Model	0.018	1.339	0.224	4472
	Proposed L2-W K-mean	**0.075**	**0.493**	0.845	**7457**
Feature Group 5	Mean Shift	0.427	0.792	0.787	304
	OPTICS	0.175	1.262	-0.335	20
	Mini Batch K-mean	**1.039**	0.797	0.804	600
	K-mean	0.898	0.879	0.897	621
	BIRCH	0.397	0.925	0.544	458
	Affinity Propagation	0.316	0.96	0.285	530
	Gaussian Mixture Model	0.653	0.937	0.382	49
	Proposed L2-W K-mean	0.965	**0.676**	**0.938**	**645**
Feature Group 6	Mean Shift	0.043	0.493	0.935	2137
	OPTICS	0.011	**0.412**	0.201	74
	Mini Batch K-mean	0.239	0.617	0.862	2790
	K-mean	0.239	0.615	0.865	2782
	BIRCH	0.093	0.943	0.862	1249
	Affinity Propagation	0.066	0.564	0.571	4195
	Gaussian Mixture Model	0.008	0.945	0.563	1253
	Proposed L2-W K-mean	**0.246**	0.469	**0.968**	**4586**

In the category of partitioning algorithms, K-mean accepts the value of k and divides the data into k number of clusters. The separation is evaluated based on the variance, minimizing the inertia and sum of squares. K-mean first selects k number of points called centroids and then clusters the samples in different groups based on Euclidian distance. Finally, K-mean determines the centroid that minimizes the inertia, i.e., the sum of square error within a cluster (13). The proposed L2-Weighted K-mean clustering algorithm also gets the optimal number of clusters from the elbow curve. The proposed model performs efficiently comparative to the traditional K-mean algorithm. Instead of getting the mean value from the original feature space, the proposed methods transform the feature space and compute the weighted mean from the transformed weighted feature space.

$$\sum_{i=0}^{n} \min_{\mu \epsilon C} (\|x_i - \mu_j\|^2) \tag{13}$$

All the clustering techniques are evaluated by using different metrics shown in Table 3. The value of the DUNN index in the proposed model is high compared to all other traditional models. The value indicates the better separability of the clusters. In FG5, i.e., the clusters based on water level, number of days, and the total depth, mini-batch K-mean performs better in terms of the DUNN index. The conventional K-mean selects initial centroids randomly, while the proposed model selects the optimal value of k by generating an elbow curve.

The value of the DUNN index in all feature groups is low, and it shows the distance between the clusters, i.e., intra-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3111112, IEEE Access

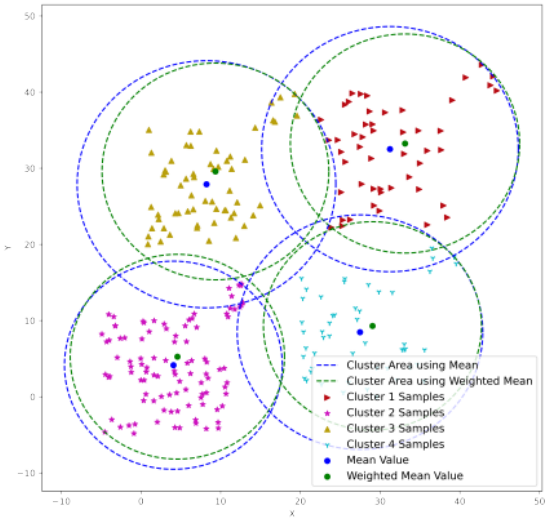


FIGURE 15: Clusters separated by L2 Weighted K- Mean Clustering Algorithm

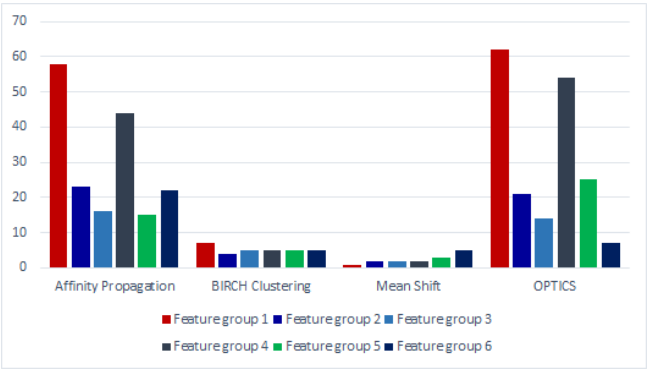


FIGURE 16: Comparison of number of cluster generated for each feature group

cluster distance. For FG1 and FG4, the value of the DUNN index indicates that the clusters are too close. In case of high cluster closure, it is required to minimize the radius of the clusters. The proposed model focuses on minimizing the radius of clusters so that if clusters reside in the close vicinity, the chances of overlapping will be reduced. At the same time, the samples will be correctly clustered.

By considering the dispersion value, i.e., CHI value, which shows the better-defined clusters, the value of CHI for each clustering technique and feature group is depicted in the bar graph in Figure 17. The high value of CHI demonstrates well-defined clusters. Except for OPTICS, a density-based algorithm that generates many tiny clusters, all other techniques established better performance in terms of CHI. The number of clusters increases the value of CHI decreases because the clusters are too close to each other. In terms of CHI index,

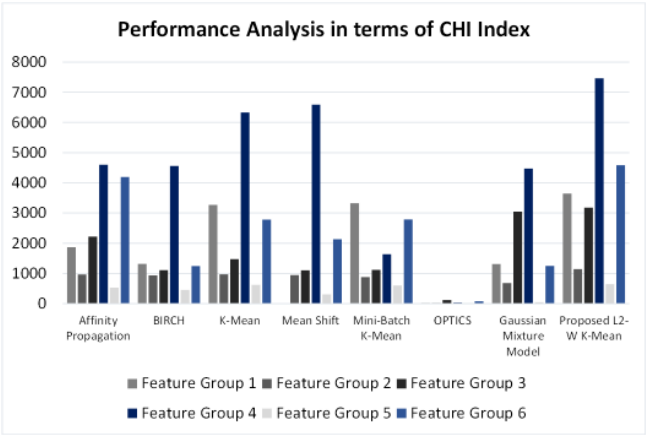


FIGURE 17: Comparison of clustering results based on CHI index

the proposed model outperforms in all feature groups. As the model focuses on the distance between the clusters, so the clusters are well separated.

In general, K-mean and mini-batch K-means are better in terms of the separation of optimal clusters from traditional clustering techniques because the elbow curve method is used to get the optimal number of clusters. Furthermore, AP and OPTICS generate a higher number of clusters than all others. The Dunn index value in the case of AP is high in some cases because the higher number of clusters in scattered areas can increase the inter-cluster distance. On the other hand, in terms of DB index and SC index, which shows the clusters' better separability, the value of OPTICS is too low because the clusters are not well separated.

The proposed model is an improved version of the K-mean. In terms of all evaluation metrics, the proposed model performs well as compared to K-mean. It is because the mean value will always shift from dense to scattered points, and the cluster radius will always be decreased. The weights are assigned to each sample using the L2 norm, which is the distance from the mean to that sample. L2-based weighted K-mean generates better and well-separated clusters as compared to all other clustering techniques.

The comparison of all the models in terms of DUNN index, DB index SC and CHI index for each feature group is shown in Figure 18. The proposed model is effective when we have closely formed clusters. However, when the clusters are closer to each other, they can overlap, and the samples can be incorrectly clustered.

V. CONCLUSIONS

Drilling and groundwater science is generating huge amounts of data from a variety of mediums and scientific experimentation. Hence, spiked attention is being given to leveraging this huge data to devise new solutions for sustainable groundwater management and overcoming hurdles endured by the drilling industry to achieve optimal drilling efficiency during borehole drilling operations. To this aim, acquiring

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3111112, IEEE Access

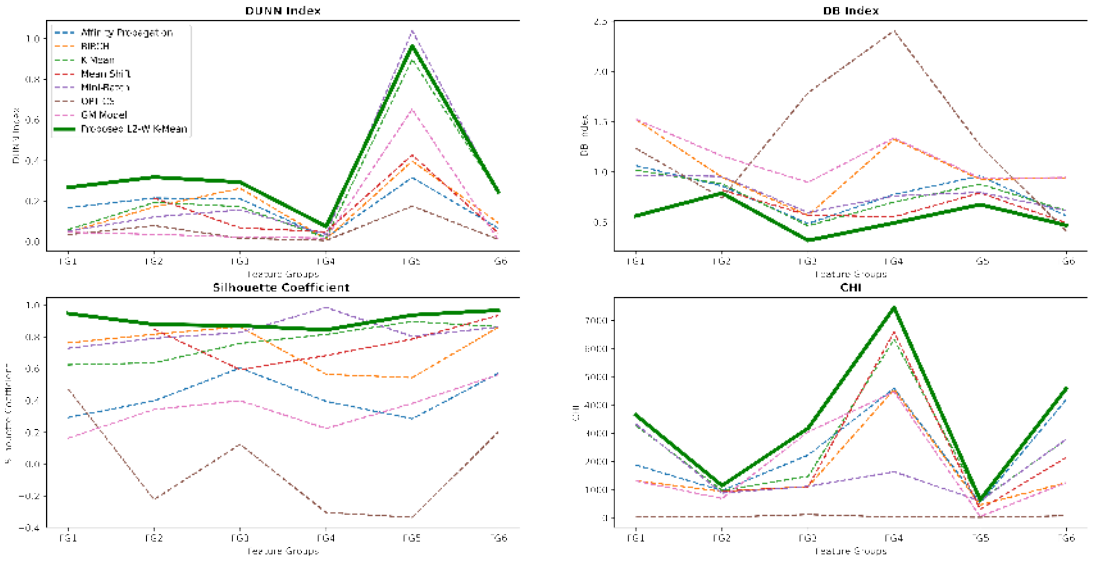


FIGURE 18: Comparison of DUNN, DB, SC and CHI indexes

prerequisite information about the drilling area and modeling groundwater processes through advanced data analytics techniques. Hence we need to analyze the area with various characteristics and attributes. To ascertain the appropriate location for digging a bore-well, some key factors should be taken into account, like the water level, number of days spent, soil color, to name a few. Hence it is imperative to analyze the given factors before initiating the drilling process; the risk factors can be minimized, such that issues like pipe stuck and high cost can be mitigated. To achieve the desired objective, we proposed a clustering model that clusters the regions based on discovering natural groupings and hidden characteristics of instances. Firstly, the data is divided into six different feature blocks, where each feature block represents the specific information. Then, by considering the specific information within these feature blocks, various clustering algorithms from four different categories are applied, and results are compared with the proposed L2 weighted K-mean clustering algorithm. The clustering techniques lead us to discover various locations based on different attributes like water level and soil color. To evaluate the effectiveness of the proposed model, unsupervised-based evaluation metrics were used. It is evident from the experimental results that the proposed model efficiently groups the data into meaningful and well-defined clusters and achieved superior performance compared to the existing ML-based clustering approaches. Furthermore, the analysis part yields useful information about soil color and the land layer's hardness level. Lastly, the data is visualized in 3D and 2D surface plots to estimate and divide different locations based on the water level, maximum depth, and soil color. The experimental results indicate that the proposed clustering algorithm's performance is fairly well and competitive to counterpart implemented clustering algorithms. The experimental results aim to improve the borehole drilling process's planning, efficiency, and management and sustainable groundwater resource management. Future work involves finding solutions for the drilling challenges based on specific use cases such as various feature sets and the development of data-driven sciences for efficient drilling operations and groundwater resource management. Furthermore, ensemble clustering techniques can be used to find more hidden patterns from different feature sets.

REFERENCES

[1] Biswas: History of hydrology - Google Scholar.
[2] Drinking Water: Equity, safety and sustainability, 2011. October 2011. Publication Title: UNICEF DATA.
[3] Manuel García-Rodríguez, Loreto Antón, and Pedro Martinez-Santos. Estimating groundwater resources in remote desert environments by coupling geographic information systems with groundwater modeling (Erg Chebbi, Morocco). Journal of Arid Environments, 110:19–29, November 2014.
[4] Chiranth Hegde and K. E. Gray. Use of machine learning and data analytics to increase drilling efficiency for nearby wells. Journal of Natural Gas Science and Engineering, 40:327–335, April 2017.
[5] Chiranth Hegde and Ken Gray. Evaluation of coupled machine learning models for drilling optimization. Journal of Natural Gas Science and Engineering, 56:397–407, August 2018.
[6] Christine I. Noshi and Jerome J. Schubert. The Role of Machine Learning in Drilling Operations; A Review. OnePetro, October 2018.
[7] Gang Rao, Guolei Wang, Xiangdong Yang, Jing Xu, and Ken Chen. Normal direction measurement and optimization with a dense three-dimensional point cloud in robotic drilling. IEEE/ASME Transactions on Mechatronics, 23(3):986–996, 2017.
[8] M Hamzaban, H Memarian, et al. Determination of relationship between drilling parameters by clustering techniques. In ISRM International Symposium-5th Asian Rock Mechanics Symposium. International Society for Rock Mechanics and Rock Engineering, 2008.

[9] Frank Nielsen. Hierarchical clustering. In Introduction to HPC with MPI for Data Science, pages 195–211. Springer, 2016.

[10] Boris Lorbeer, Ana Kosareva, Bersant Deva, Dženan Softić, Peter Ruppel, and Axel Küpper. Variations on the clustering algorithm birch. Big data research, 11:44–53, 2018.

[11] Christian Borgelt and Rudolf Kruse. Agglomerative fuzzy clustering. In International Conference on Soft Methods in Probability and Statistics, pages 69–77. Springer, 2016.

[12] Jianyun Lu and Qingsheng Zhu. An effective algorithm based on density clustering framework. Ieee Access, 5:4991–5000, 2017.

[13] Yinghua Lv, Tinghuai Ma, Meili Tang, Jie Cao, Yuan Tian, Abdullah Al-Dhelaan, and Mznah Al-Rodhaan. An efficient and scalable density-based clustering algorithm for datasets with complex structures. Neurocomputing, 171:9–22, 2016.

[14] Paul D McNicholas. Model-based clustering. Journal of Classification, 33(3):331–373, 2016.

[15] Cinzia Viroli and Geoffrey J McLachlan. Deep gaussian mixture models. Statistics and Computing, 29(1):43–51, 2019.

[16] Yawei Zhao, Yuewei Ming, Xinwang Liu, En Zhu, Kaikai Zhao, and Jianping Yin. Large-scale k-means clustering via variance reduction. Neurocomputing, 307:184–194, 2018.

[17] Debra Perrone and Scott Jasechko. Deeper well drilling an unsustainable stopgap to groundwater depletion. Nature Sustainability, 2(8):773–782, 2019.

[18] N Hassan, H Zabidi, KS Ariffin, and M Trisugiwo. Effect of drilling speed of probe drilling data on the clustering of rock strength at natm-4, hulu langat, selangor. Procedia Chemistry, 19:737–742, 2016.

[19] Rahman Ashena, Minou Rabiei, Vamegh Rasouli, Amir H Mohammadi, et al. Optimization of drilling parameters using an innovative ga-ps artificial intelligence model. In SPE Asia Pacific Oil & Gas Conference and Exhibition. Society of Petroleum Engineers, 2020.

[20] Chandan Guria, Kiran K Goli, and Akhilendra K Pathak. Multi-objective optimization of oil well drilling using elitist non-dominated sorting genetic algorithm. Petroleum Science, 11(1):97–110, 2014.

[21] Huadong Guo. Big earth data: A new frontier in earth and information sciences. Big Earth Data, 1(1-2):4–20, 2017.

[22] Amir Gandomi and Murtaza Haider. Beyond the hype: Big data concepts, methods, and analytics. International journal of information management, 35(2):137–144, 2015.

[23] Amit Mehta. Tapping the value from big data analytics. Journal of Petroleum Technology, 68(12):40–41, 2016.

[24] Jina Jeong and Eungyu Park. Comparative application of various machine learning techniques for lithology predictions. Journal of Soil and Groundwater Environment, 21(3):21–34, 2016.

[25] Kevin P Murphy. Machine learning: a probabilistic perspective. MIT press, 2012.

[26] Sean D Kristjansson, Adam Neudfeldt, Stephen W Lai, Julian Wang, Dean Tremaine, et al. Use of historic data to improve drilling efficiency: a pattern recognition method and trial results. In IADC/SPE Drilling Conference and Exhibition. Society of Petroleum Engineers, 2016.

[27] AS Shihab and A Hashim. Cluster analysis classification of groundwater quality in wells within and around mosul city. Iraq. J. Environ. Hydrol, 14, 2006.

[28] Naeem Iqbal, Atif Rizwan, Anam Nawaz Khan, Rashid Ahmad, Bong Wan Kim, Kwangsoo Kim, and Do-Hyeun Kim. Boreholes Data Analysis Architecture Based on Clustering and Prediction Models for Enhancing Underground Safety Verification. IEEE Access, 9:78428–78451, 2021. Conference Name: IEEE Access.

[29] Chris Carpenter. Stuck-pipe prediction with automated real-time modeling and data analysis. Journal of Petroleum Technology, 68(06):72–73, 2016.

[30] Cüneyt Güler and Geoffrey D Thyne. Delineation of hydrochemical facies distribution in a regional groundwater system by means of fuzzy c-means clustering. Water Resources Research, 40(12), 2004.

[31] Tina Helstrup, Niels Oluf Jørgensen, and Bruce Banoeng-Yakubo. Investigation of hydrochemical characteristics of groundwater from the cretaceous-eocene limestone aquifer in southern ghana and southern togo using hierarchical cluster analysis. Hydrogeology Journal, 15(5):977–989, 2007.

[32] Cüneyt Güler, Geoffrey D Thyne, John E McCray, and Keith A Turner. Evaluation of graphical and multivariate statistical methods for classification of water chemistry data. Hydrogeology journal, 10(4):455–474, 2002.

[33] Andrew R Webb. Statistical pattern recognition. John Wiley & Sons, 2003.

[34] Lilik Eko Widodo, Tedy Agung Cahyadi, Sudarto Notosiswoyo, and Eman Widijanto. Application of clustering system to analyze geological, geotechnical and hydrogeological data base according to hc-system approach. 2017.

[35] Moslem Fatehi and Hooshang H Asadi. Application of semi-supervised fuzzy c-means method in clustering multivariate geochemical data, a case study from the dalli cu-au porphyry deposit in central iran. Ore Geology Reviews, 81:245–255, 2017.

[36] Clemens Reimann, Peter Filzmoser, Robert Garrett, and Rudolf Dutter. Statistical data analysis explained: applied environmental statistics with R. John Wiley & Sons, 2011.

[37] Atif Rizwan, Naeem Iqbal, Rashid Ahmad, and Do-Hyeun Kim. WR-SVM Model Based on the Margin Radius Approach for Solving the Minimum Enclosing Ball Problem in Support Vector Machine Classification. Applied Sciences, 11(10):4657, January 2021. Number: 10 Publisher: Multidisciplinary Digital Publishing Institute.

[38] Horst Bunke and Alberto Sanfeliu. Syntactic and structural pattern recognition: theory and applications, volume 7. World Scientific, 1990.

[39] Dong Huang, Chang-Dong Wang, Hongxing Peng, Jianhuang Lai, and Chee-Keong Kwoh. Enhanced Ensemble Clustering via Fast Propagation of Cluster-Wise Similarities. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 51(1):508–520, January 2021. Conference Name: IEEE Transactions on Systems, Man, and Cybernetics: Systems.

[40] Dong Huang, Chang-Dong Wang, Jian-Huang Lai, and Chee-Keong Kwoh. Toward Multidiversified Ensemble Clustering of High-Dimensional Data: From Subspaces to Metrics and Beyond. IEEE Transactions on Cybernetics, pages 1–14, 2021. Conference Name: IEEE Transactions on Cybernetics.

[41] Carol Friedman. System and method for language extraction and encoding utilizing the parsing of text data in accordance with domain parameters, January 30 2001. US Patent 6,182,029.

[42] Shichao Zhang. Nearest neighbor selection for iteratively knn imputation. Journal of Systems and Software, 85(11):2541–2552, 2012.

[43] Pritesh Vora, Bhavesh Oza, et al. A survey on k-mean clustering and particle swarm optimization. International Journal of Science and Modern Engineering, 1(3):24–26, 2013.

[44] Carl Edward Rasmussen. The Infinite Gaussian Mixture Model. page 7.

[45] Md. Mostofa Ali Patwary, Diana Palsetia, Ankit Agrawal, Wei-keng Liao, Fredrik Manne, and Alok Choudhary. Scalable parallel OPTICS data clustering using graph algorithmic techniques. November 2013.

[46] Christopher R Palmer and Christos Faloutsos. Density Biased Sampling: An Improved Method for Data Mining and Clustering. page 11.

[47] Jeffrey S. Vitter. Random sampling with a reservoir. ACM Transactions on Mathematical Software, 11(1):37–57, March 1985.

[48] Kim-Hung Li. Reservoir-sampling algorithms of time complexity $O(n(1 + \log(N/n)))$. ACM Transactions on Mathematical Software, 20(4):481–493, December 1994.

[49] Frank Olken, Doron Rotem, and Ping Xu. Random sampling from hash files. ACM SIGMOD Record, 19(2):375–386, May 1990.

[50] Fang-Xiang Wu. Genetic weighted k-means algorithm for clustering large-scale gene expression data. BMC Bioinformatics, 9(S6):S12, May 2008.

ATIF RIZWAN is currently pursuing a Ph.D. in the Department of Computer Engineering at Jeju National University, Republic of Korea. He received his M.S. in Computer Science from COMSATS University Islamabad, Attock Campus, Punjab, Pakistan, in 2020. He has also completed his Master of Computer science (16 years) from the COMSATS University Islamabad, Attock Campus. He has good industry experience in software development and testing. His research work focused on Machine Learning, Data and Web Mining, Analysis of optimization of Core Algorithms, and IoT-based Applications.

NAEEM IQBAL is currently pursuing a Ph.D. in the Department of Computer Engineering at Jeju National University, Republic of Korea. He received his M.S in Computer Science from COMSATS University Islamabad, Attock Campus, Punjab, Pakistan, in 2019. He did his B.S in Computer Science from the COMSATS University Islamabad, Attock Campus, Pakistan. He has professional experience in the software development industry and in academics as well. His research work mainly focused on Machine Learning, Big Data, AI-based Intelligent Systems, Analysis of Optimization Algorithms, IoT and Blockchain-based Applications.

ANAM NAWAZ KHAN is currently pursuing a Ph.D. at the Department of Computer Engineering Jeju National University, Republic of Korea. She received an M.S degree in Computer Science from COMSATS University Islamabad, Attock Campus, Pakistan in 2019. She did B.S in Computer Science from the COMSATS University Islamabad, Attock Campus, in 2016. Her research work mainly focused on machine learning applications in smart environments, analysis of prediction and optimization algorithms, big data and IoT-based Applications.

RASHID AHMAD received the B.S. degree from the University of Malakand, Pakistan, in 2007, the M.S. degree in Computer Science from the National University of Computer and Emerging Sciences (NUCES), Islamabad, Pakistan, in 2009, and the Ph.D. degree in computer engineering from Jeju National University, Republic of Korea, in 2015. Since 2016, he has been with COMSATS University Islamabad, Attock Campus, Pakistan, where he is currently an Assistant Professor with the Department of Computer Science. His research work is focused on the application of prediction and optimization algorithms to build IoT-based solutions. His research interests mainly focused on Machine Learning, Data Mining, related applications.

DOHYEUN KIM received the B.S. degree in electronics engineering from Kyungpook National University, South Korea, in 1988, and the M.S. and Ph.D. degrees in information telecommunication from Kyungpook National University, the Republic of Korea , in 1990 and 2000, respectively. He was with the Agency of Defense Development (ADD), from 1990 to 1995. Since 2004, he has been with Jeju National University, the Republic of Korea , where he is currently a Professor with the Department of Computer Engineering. From 2008 to 2009, he was a Visiting Researcher with the Queensland University of Technology, Australia. His research interests include sensor networks, M2M/IOT, energy optimization and prediction, intelligent service, and mobile computing.